SHORT PROTOCOLS IN MOLECULAR BIOLOGY

Third Edition

A Compendium of Methods from
Current Protocols in Molecular Biology

EDITORIAL BOARD

Frederick M. Ausubel
Massachusetts General Hospital & Harvard Medical School

Roger Brent
Massachusetts General Hospital & Harvard Medical School

Robert E. Kingston
Massachusetts General Hospital & Harvard Medical School

David D. Moore
Baylor College of Medicine

J.G. Seidman
Harvard Medical School

John A. Smith
University of Alabama at Birmingham

Kevin Struhl
Harvard Medical School

WILEY

Published by John Wiley & Sons, Inc.

New York • Chichester • Weinheim • Brisbane • Singapore • Toronto

DNA S quencing

7

OVERVIEW OF DNA SEQUENCING METHODS

For many recombinant DNA experiments, knowledge of a DNA sequence is a prerequisite for its further manipulation. DNA sequencing followed by computer-assisted searching for restriction endonuclease cleavage sites is often the fastest method for obtaining a detailed restriction map (*UNITS 3.1-3.3*). This information is particularly useful when vectors designed to overexpress proteins or to generate protein fusions are utilized for subcloning a gene of interest (Chapter 16). Computer-assisted identification of protein-coding regions (open reading frames or ORFs) within the DNA sequence, followed by computer-assisted similarity searches of DNA and protein data bases, can lead to important insights about the function and structure of a cloned gene and its product (*UNIT 7.7*). In addition, the DNA sequence is a prerequisite for a detailed analysis of the 5′ and 3′ noncoding regulatory regions of a gene. DNA sequence information is essential for site-directed mutagenesis (*UNIT 8.1*). Small amounts of DNA sequence information (sequence tagged sites, or STS, or expressed sequence tags, or EST) are the basis of methods for mapping and ordering large DNA segments cloned into yeast artificial chromosomes or cosmids.

DNA sequencing techniques are based on electrophoretic procedures using high-resolution denaturing polyacrylamide (sequencing) gels. These so-called sequencing gels are capable of resolving single-stranded oligonucleotides up to 500 bases in length which differ in size by a single deoxynucleotide. In practice, for a given region to be sequenced, a set of labeled, single-stranded oligonucleotides is generated, the members of which have one fixed end and which differ at the other end by each successive deoxynucleotide in the sequence. The key to determining the sequence of deoxynucleotides is to generate, in four separate enzymatic or chemical reactions, all oligonucleotides that terminate at the variable end in A, T, G, or C. The oligonucleotide products of the four reactions are then resolved on adjacent lanes of a sequencing gel. Because all possible oligodeoxynucleotides are represented among the four lanes, the DNA sequence can be read directly from the four "ladders" of oligonucleotides as shown in Figure 7.0.1.

The practical limit on the amount of information that can be obtained from a set of sequencing reactions is the resolution of the sequencing gel (see *UNIT 7.6* for protocols on setting up and running sequencing gels). Current technology allows ~300 nucleotides of sequence information to be reliably obtained in one set of sequencing reactions, although more information (up to 500 nucleotides) is often obtained. Thus, if the region of DNA to be sequenced is <300 nucleotides, a single cloning into the appropriate vector is all that is usually necessary to produce a recombinant molecule that can easily be sequenced.

For a larger region of DNA, it is generally necessary to break a large fragment into smaller ones that are then individually sequenced. This can be done in a random or an ordered fashion. *UNIT 7.1* contains a discussion of strategies for sequencing large regions of DNA. Two protocols for subdividing large regions of DNA are provided in *UNIT 7.2*. These protocols are used to create a set of ordered, or nested, deletions for DNA sequencing using exonuclease III or nuclease *Bal*31.

The two methods that are widely used to determine DNA sequences, the enzymatic dideoxy method and the chemical method, differ primarily in the technique used to generate the ladder of oligonucleotides. In the enzymatic dideoxy sequencing method, a DNA polymerase is utilized to synthesize a labeled, complementary copy

sequence from an open reading frame (ORF); many restriction mapping programs also include a translation feature. Translating the sequence in this fashion provides a simple check for deletions and insertions. Some software is able to take into account variant genetic codes; for instance, DNA Strider allows the user to select from a list of variant codes, and the GCG package allows the user to reset individual codons. Although this technique will detect only a subset of possible errors, it can often quickly identify common problems such as simple typing mistakes or miscounting of the number of the same nucleotides in a run of identical nucleotides.

Most DNA sequence packages allow the user to specify a range and reading frame to be used in translating a DNA sequence. DNA Strider (Table 7.7.1) has a useful feature that hunts for ORFs in a DNA sequence and highlights the putative coding sequence, thus doing all the necessary work.

Detecting Overlap with Other Sequenced Fragments

With increasing worldwide interest in genome sequencing projects, sequence assembly packages now provide very effective automatic DNA sequence assembly, connecting shorter pieces of DNA to build the longest continuous sequence possible. Once the sequences to be assembled are identified, they are compared, overlaps identified, and contiguous sequences (contigs) constructed. Typically, parameters are available to allow adjustment of the alignment process. A good commercial assembly program is LaserGene, available from DNAStar for both Macintosh and IBM-compatible computers (Table 7.7.1). Once a sequence contig has been assembled, the LaserGene program creates a graphic overview of the sequencing project, highlighting regions that need further verification. Another commercial program is Sequencher (Gene Codes). This program provides expanded features for the assembly, processing, and editing of DNA sequences determined with the ABI sequencer; however, it is only available for the Macintosh computer.

If a sophisticated assembly program is not available, contigs can be constructed "by hand" using a comparison program and a multiple sequence editor. Both the input DNA sequence and its complement should be considered in assembling contigs. When using a program that does not automatically consider the complement sequence, conduct a separate search of that sequence.

When conducting sequence comparisons to identify overlaps, it is important to take into account the fact that different programs take slightly different approaches to this process. Some comparison programs that use the Needleman and Wunsch algorithm (e.g., the GCG GAP program) are designed to find the maximum number of matches between two sequences over the entire length of the two sequences, with the minimum number of gaps. This type of program is best suited for aligning two sequences along their entire lengths. If the two sequences differ greatly in size, however, the result may not satisfactorily represent the similarity between them. For identifying regions of similarity between two pieces of DNA that are largely different in sequence or of significantly different lengths, programs employing algorithms such as those described by Smith and Waterman or Wilbur and Lipman—e.g., the GCG BESTFIT program—are more suitable. These algorithms do not try to align the two sequences being compared in their entirety, but instead search for short matches within the sequences.

Editing a Contig and Verifying the Sequence

Generally, software packages that provide sequence assembly programs include multiple sequence editors that display the individual sequences of the aligned contig together one on top of the other, one sequence per line (see Fig. 7.7.2), and can

BEST AVAILABLE COPY